Case description from the project “FAIRify Humanities” of the National Data management Forum 2019.

2019-12-20

***Title**:* FAIRifying Danish threatening messages

***Name and affiliation**:* Tanya Karoli Christensen, UCPH

***Research area**:* The research project that has generated this data set is called “Understanding Threats: Language and Genre” and is placed within the field of forensic linguistics. Comparatively little is known about grammatical and stylistic traits of threatening language. Extending results for English data to Danish, a corpus of authentic threatening letters is analyzed qualitatively and quantatively to discern the functions and distributional patterns of linguistic features related to the main characteristics of a verbal threat, namely expressions of futurity, intended harm and author responsibility. The project is funded by the Danish Carlsberg Foundation.

Data set(s): 120 threatening messages in Danish, all included as facsimile in two previously published books compiled by journalists Robin Engelhardt and Christian Lund. No GDPR restrictions.

***What efforts was carried out to make your data more FAIR**?*

In general: All messages were first transcribed in a text editor. Each text was then converted into xml-format and provided with metadata conforming to TEI 5 standards.

More “Findable”: The data set will be uploaded to a corpus search tool KORP and will be provided as a downloadable dataset with added metadata, which is made human searchable and machine harvestable via the OAI-PMH protocol. The dataset can be downloaded from this PID: http://hdl.handle.net/20.500.12115/40.

More “Accessible”: We have obtained permission by the publisher for online publication for five years, from 2020-2025, with no restrictions on how researchers and students use the annotated data. This period is extendable upon application.

More “Interoperable”: The individual texts were xml-coded according to TEI P5 standard and were PoS- and lemma-tagged according to a (slightly modified) PAROLE standard. The TEI P5 standard is widely used in research communities using text material and is well documented.

More “Re-usable”: As stated in the metadata, this dataset is this dataset is licensed to non-commercial use, meaning that anybody can use the data, except for commercial purposes.

What was the biggest challenges to make your data more FAIR? Only a subset of all the research data related to this project can ever be really FAIR since the majority of our data are highly sensitive and we are not allowed to make them publicly available. But even the set that can be made available for others, required some negotiation with the publisher who, unsurprisingly, is careful not to give up on their publication rights. Since I ultimately would like users to be able to see a facsimile of each text in addition to the linguistically annotated texts, a full online publication of the corpus would make later print publications less profitable. Another obstacle is getting funding for student assistants to add metadata and to correct automatic PoS- and lemma-tagging, but this goes for all corpus establishing projects.